International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-7, Issue-6, November 2019
https://doi.org/10.21276/ijircst.2019.7.6.2
www.ijircst.org

Twitter Data Classification by Applying and Comparing Multiple Machine Learning Techniques

Ananya Sarker, Md. Shahid Uz Zaman, Md. Azmain Yakin Srizon

***Abstract*- Having an average of five hundred million tweets sent out per day, twitter has become one of the largest platforms of data analysis for the researchers. Previously, various researches have been conducted on twitter data i.e., sentimental analysis. However, not much research has been done to classify the tweets in terms of categories so that tweets can be distributed as per user preferences. In this research we started by creating four broad categories: politics, sports, crime and natural. After that, we applied different machine learning techniques (Random Forest, K-Nearest Neighbors, Naïve Bayes, Logistic Regression, Decision Tree and Support Vector Machine) to classify the twitter data. Finally, we compared the results in terms of sensitivity, specificity, precision, false positive rate and accuracy. We found that Support Vector Machine (SVM) produced the best results in terms of sensitivity, specificity, precision, false positive rate and accuracy. Hence, we concluded that a machine learning approach (Support Vector Machine) can certainly be used to classify twitter data. Constructed dataset, all the programs, figures and snippets can be found at https://github.com/ananyasarkertonu/Twitter-Dataset**

***Keywords*- Classification, Machine Learning, Social Media, Twitter Data.**

I. INTRODUCTION

Social media is a web-based technology to smooth social interaction between large groups of people through some type of network. One of the most used micro blogging websites is twitter. Different social networking sites:

Twitter, LinkedIn, Google+ are becoming more popular as they allow people to share and express their opinion about topics, have a discussion with different communities, or post messages across the world. There are many works have been done in the field of sentiment analysis of twitter data. Sentiment analysis or classification of twitter data is helpful to analyze the electronic text. Some of the opinions of twitter data are highly unstructured and are either positive or negative, or neutral in some cases [1].

At the present time people are trying to develop a system that can identify and classify the viewpoint of different electronic texts such as tweets. There are different procedures to recognize the sentiments from twitter posts and predict online customer's favorites, sometimes it is valuable for economic and market research. Classified tweets about positive, negative or neutral indicates the opinions of people on product that help users to buy the best product [2]. Another research about sentiment analysis on twitter data is that various machine learning approaches in a hybrid manner gives more accurate result instead of using these machine learning approaches in isolation. K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) are in a hybrid manner gives more accuracy and classified tweets into positive, negative and neutral sentiments [3]. Natural language processing supports to extract information about sentiment or opinion from a text. Twitter sentiment analysis is an application of the text classification problem. The main approach of this type of research is to connect with Twitter and collect tweets that contain a particular keyword and calculate the polarity of the tweets as positive or negative or neutral. Another approach of sentiment classification is in different level: sentence level, document level or phrase level [4] [5]. Some research about real-time tweet summarization of scheduled sub-events of different type of game like FIFA world cup-2015, IPL-2015 etc. this type of research use twitter social media data to analyze sentiment of games, feelings of fans, evaluation polarity and so on [6]. A large amount of data from social media like twitter is very difficult to handle as there are many unwanted information. Bharati S. Kannolli and Prabhu R. Bevinmarad implement unsupervised methods for sentiment analysis of cricket match. They only predict the outcome of a cricket match as a winning team or losing team based on the tweets posted by their fans or user [7].

So, there are many researches about sentiment classification on twitter data with different topic like electronic product, movie reviews, cricket tournament,

Manuscript received November 20, 2019

Ananya Sarker, Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh, +8801739166530, (e-mail: ananya.ruet@gamil.com)

Md. Shahid Uz Zaman, Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh, +8801713228537

Md. Azmain Yakin Srizon, Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh, +8801790187189

football match etc. But in this research, we started by creating four extensive categories: politics, sports, crime and natural. Then we applied six machine learning techniques (Random Forest, K-Nearest Neighbors, Naïve Bayes, Logistic Regression, Decision Tree and Support Vector Machine) to classify the twitter data and finally we compared these classification techniques.

II. PROPOSED ARCHITECTURE

In this research the work started by collecting twitter data and splitting these into training set and testing set. Then preprocessing these data so that these can be fit for feature extraction, then classification techniques applied on the preprocessed dataset. Six classification techniques were applied and compared each other with accuracies. Figure 1 shows the steps of work flowchart.

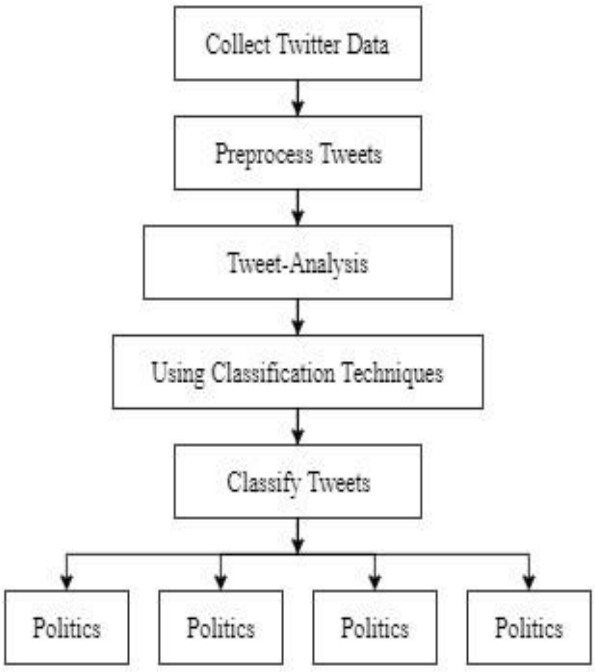

Fig 1: Work procedure to classify tweets

The classification techniques applied on the twitter dataset which are acquired from the developer account. Table 1 shows the data set description.

Table 1: The description of data set

Criteria	Description
Dataset Name	Twitter Data
Dataset URL	https://github.com/ananyasarkertonu /Twitter-Dataset
File Type	CSV format
Tweets	Train set 603, Test set 100
Classification Type	Politics, Sports, Natural, Crime

III. METHODOLOGY

The work started by collecting twitter data and splitting dataset into training and testing set. The training set contains 603 sample tweets and test set contain 100 sample tweets. These tweets were not suitable for feature extraction as these contain some unnecessary feature like hash tags, username, special character, URL's etc. so next step was data preprocessing. Seven classification techniques Random Forest (RF), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT) and Support Vector Machine (SVM) were applied on these preprocessed datasets.

A. Dataset Description

For collecting twitter dataset first have a developer account. Twitter provide a platform to access data from account and anyone can use this for own purpose. For this create a new Twitter application to get OAuth credentials and API access. After creating API access, we collect customer key, customer secret key, access token key and access secret key and with the help of these keys we collect the data set. As our purpose to analyze general tweets and classify these tweets into politics, sports, crime and natural so apply a filter to this particular topic to get desired tweets with limited time interval. Thus, tweets are imported from twitter. The constructed dataset can be found at https://github.com/ananyasarkertonu/Twitter-Dataset.

B. Dataset Preprocessing

Several functions of NLTK were used for data preprocessing. In the preprocessing stage first extracted main information from tweets and removed all unnecessary contents. Table 2 shows some of the unnecessary contents which were removed from the main tweets.

Table 2: Table of some removed contents from original tweet

Contents	Actions
Punctuation	Removed
Uppercase Character	Lowercase all contents
All word	converted into simple form
Empty space	Removed
Number	Removed
#word	Removed #

C. Classification Techniques

Here we describe Random Forest (RF), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT) and Support Vector Machine (SVM) classification techniques. These classification techniques were used for twitter text classification.

1) Random Forest (RF) Classifier

Random Forest is a classifier consists of many tree-based structures. It runs efficiently on a large dataset and can handle thousands of input variables. It also gives estimates of what variables are important in the classification. This method can also handle the over fitting of data points. For a dataset, D, with N instances and A attributes, the general

procedure to build a Random Forest classifier is as follows. For each time of construction, a candidate Decision Tree, a subset of the dataset D, d, is sampled with replacement as the training dataset. In decision tree, for every node an arbitrary subset of the attributes A, a, is selected as the candidate attributes to split the node. By building K Decision Trees in this way, a Random Forest classifier is built [8]. In this work the number of estimators was set to 100, max depth was set to 10, the random set was 0, the criterion was set to 'Entropy' and all other parameters were set to default.

2) K-Nearest Neighbors (KNN) Classifier

K- Nearest Neighbors is one of the simplest algorithms in machine learning which is used for both classification and regression problems. It is a non-parametric algorithm and based on feature similarity. This classifier is a lazy learner because it does not use the training data for generalization. In this work, the number of neighbors was set to 5, weights were uniformly distributed, leaf size was set to 30 and all other parameters were set to default.

3) Naïve Bayes (NB) Classifier

Naïve Bayes' algorithm is a probabilistic classifier. This classifier based on Bayes' theorem. This classifier is very useful for large datasets and widely used because of its simplicity. A Naïve Bayes classifier consists of two components: quantitative and qualitative. The quantitative components of Naïve Bayes classifier can be represented in form of network parameters called conditional table while the qualitative components of Naïve Bayes classifier can be represented in form of network structure [9]. The probability equation is given below

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} \quad (1)$$

Here P(A/B) is the Conditional probability of occurrence of event A given the event B,

P(B/A) is the likelihood which is the probability of predictor given class,

P(A) is the prior probability of class and

P(B) is the prior Probability of the predictor

In this work, all the parameters of the Naïve Bayes Classifier were set to default.

4) Logistic Regression (LR) Classifier

Logistic Regression is a classification algorithm based on the idea of probability. If the number of input passes through the logistic function then the function gives result with the probability score between 0 to 1. This algorithm works on more complex cost function which is known as sigmoid function as well as logistic function. The sigmoid function is defined as follows:

$$f(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}} \quad (1)$$

For this work, in Logistic Regression algorithm penalty was set to 12, max iteration was 100, and all other parameters were set to default to get more accurate results.

5) Decision Tree (DT) Classifier

Decision Tree is a simple and popular classification technique for classifying data. Decision tree is a tree-based structure that includes a root, branches, and some leaf nodes. The interior node of a decision tree represents the test on an attribute and each branch denotes the outcome of a test, and each leaf node holds a class label. The top node in the decision tree is called the root node. Decision tree consists of two phases, at first tree construction is done by assigning all the training examples are at the root node then partition recursively based on selected attributes. Then tree pruning identifies and remove branches that reflect noise or outliers. This classification algorithm also handles the over-fitting problem using tree pruning process. In this work, all the parameters of the decision tree algorithm were set to default.

6) Support Vector Machine (SVM) Classifier

Support Vector Machine is a linear model for classification and regression problems. This classifier can solve many linear and non-linear practical problems. It creates a hyper-plane that converts the set of data into classes. SVM is also used for multi-class classification problem. Fig 2 shows the hyper-plane used in the classification of two classes.

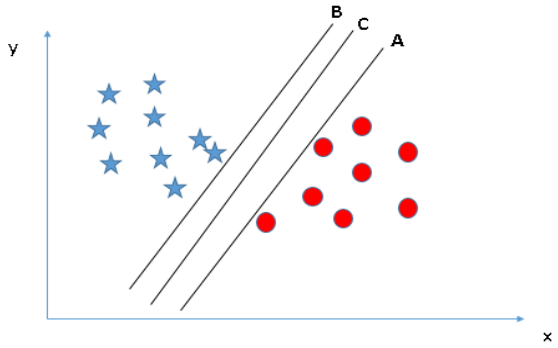

Fig 2: Hyper-plane used for classification in SVM

While implementing Support Vector Machine algorithm in this research, multiclass was set to 'crammer_singer', random state was set to none and all other parameters were set to default and this algorithm gives more accuracy.

IV. EXPERIMENTAL ANALYSIS

The procedure started by splitting the whole dataset into train and test sets where the train set contains 603 tweet samples and the test set contains 100 tweet samples. Each set contains the tweet label (sports, politics, crime, and natural) on the first column and twitter data on the second column. The twitter data however had some redundant information attached with the tweet. Beautiful Soup was used to remove the redundant data. Word tokenizer was used to extract the tokens from the string of characters. The next step was to create the dictionaries. But before the construction of the dictionaries, the word which does not add much meaning to the sentence was removed from the tokens with the help of the stopwords function. The language was set to English while removing the redundant words. After that stemming process was executed to reduce morphological variants of a root or base word. Finally, four dictionaries were created as

the problem domain had four class labels. Each of the dictionaries contained the tokens of the tweets of the corresponding classes. For example, if a tweet had a class label of sports, all the tokens generated from that tweet were stored in the sports dictionary. Some of the words were present in multiple dictionaries. These words could cause accuracy loss because of the lack of relevance to a particular class. Hence, these words were eliminated from the dictionaries as well. After the construction of the dictionaries, the feature vectors were created for applying classification methods. Scikit-learn was used to produce the classifier for the twitter data classification. Classification techniques used in this research were Random Forest (RF), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT) and Support Vector Machine (SVM). While using Random Forest classifier, number of estimators was set to 100, max depth was set to 10, random set was 0, criterion was set to 'Entropy' and all others parameters were set to default. Table-3 shows the sensitivity, specificity, precision and false positive rate of Random Forest algorithm.

Table 3: Sensitivity, specificity, precision and false positive rate calculation for Random Forest algorithm

Class Name	Sensitivity	Specificity	Precision	False Positive Rate
Politics	0.00	1.00	0.00	0.00
Sports	0.96	0.04	0.45	0.96
Natural	0.03	0.97	0.33	0.03
Crime	0.00	1.00	0.00	0.00
Average	**0.25**	**0.75**	**0.20**	**0.25**

Table 4: Sensitivity, specificity, precision and false positive rate calculation for K-Nearest Neighbors algorithm

Class Name	Sensitivity	Specificity	Precision	False Positive Rate
Politics	0.70	0.88	0.58	0.13
Sports	0.98	0.60	0.67	0.40
Natural	0.19	1.00	1.00	0.00
Crime	0.00	1.00	0.00	0.00
Average	**0.47**	**0.87**	**0.56**	**0.13**

Table 5: Sensitivity, specificity, precision and false positive rate calculation for Naïve Bayes algorithm

Class Name	Sensitivity	Specificity	Precision	False Positive Rate
Politics	0.55	1.00	1.00	0.00
Sports	0.76	0.96	0.95	0.04
Natural	1.00	0.80	0.69	0.20
Crime	0.75	1.00	1.00	0.00
Average	**0.77**	**0.94**	**0.91**	**0.06**

Table 6: Sensitivity, specificity, precision and false positive rate calculation for Logistic Regression algorithm

Class Name	Sensitivity	Specificity	Precision	False Positive Rate
Politics	0.65	1.00	1.00	0.00
Sports	0.82	0.95	0.93	0.05
Natural	0.97	0.87	0.77	0.13
Crime	0.75	1.00	1.00	0.00
Average	**0.80**	**0.96**	**0.92**	**0.05**

While implementing K-Nearest Neighbors algorithm, number of neighbors was set to 5, weights were uniformly distributed, leaf size was set to 30 and all other parameters were set to default. Table-4 shows the sensitivity, specificity, precision and false positive rate of K-Nearest Neighbors algorithm. While implementing Naïve Bayes algorithm, all the parameters were set to default. Table-5 shows the sensitivity, specificity, precision and false positive rate of Naïve Bayes algorithm. While implementing Logistic Regression algorithm, penalty was set to 12, max iteration was 100, and all other parameters were set to default. Table-6 shows the sensitivity, specificity, precision and false positive rate of Logistic Regression algorithm.

While implementing Decision Tree algorithm, all the parameters were set to default. Table-7 shows the sensitivity, specificity, precision and false positive rate of Decision Tree algorithm.

Table 7: Sensitivity, specificity, precision and false positive rate calculation for Decision Tree algorithm

Class Name	Sensitivity	Specificity	Precision	False Positive Rate
Politics	0.70	1.00	1.00	0.00
Sports	0.82	0.96	0.95	0.04
Natural	0.97	0.84	0.76	0.16
Crime	0.75	1.00	1.00	0.00
Average	**0.81**	**0.95**	**0.93**	**0.05**

While implementing Support Vector Machine algorithm, multiclass was set to 'crammer_singer', random state was set to none and all other parameters were set to default. Table-8 shows the sensitivity, specificity, precision and false positive rate of Support Vector Machine algorithm.

Table 8: Sensitivity, specificity, precision and false positive rate calculation for Support Vector Machine algorithm

Class Name	Sensitivity	Specificity	Precision	False Positive Rate
Politics	0.70	1.00	1.00	0.00
Sports	0.82	0.96	0.95	0.04
Natural	0.97	0.89	0.77	0.11
Crime	0.75	1.00	1.00	0.00
Average	**0.81**	**0.96**	**0.93**	**0.04**

Table-9 shows the comparison result of six algorithms in

terms of sensitivity, specificity, precision, false positive rate and accuracy.

Table 9: Comparison of sensitivity, specificity, precision, false positive rate and accuracy among all six algorithms

	RF	KNN	NB	LR	DT	SVM
Sensitivity	0.25	0.47	0.77	0.80	**0.81**	**0.81**
Specificity	0.75	0.87	0.94	**0.96**	0.95	**0.96**
Precision	0.20	0.56	0.91	0.92	**0.93**	**0.93**
FPR	0.25	0.13	0.06	0.05	0.05	**0.04**
Accuracy	0.44	0.64	0.80	0.83	**0.84**	**0.84**

Fig 3, Fig 4, Fig 5, Fig 6, Fig 7 shows the comparison bar graph of sensitivity, specificity, precision, false positive rate and accuracy of Random Forest (RF), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT) and Support Vector Machine (SVM) respectively.

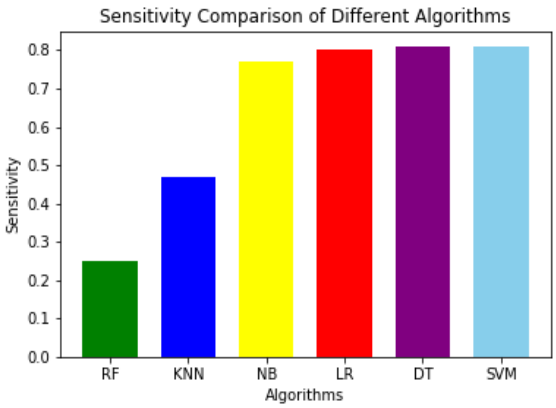

Fig 3: Comparison of sensitivity among all algorithms.

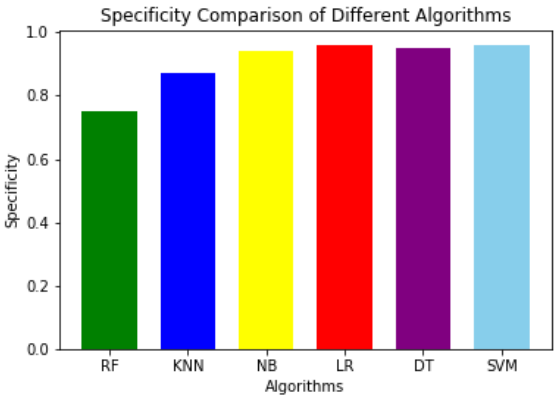

Fig 4: Comparison of specificity among all algorithms.

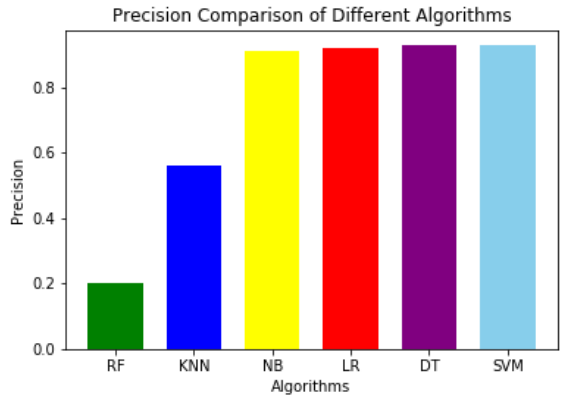

Fig 5: Comparison of precision among all algorithms.

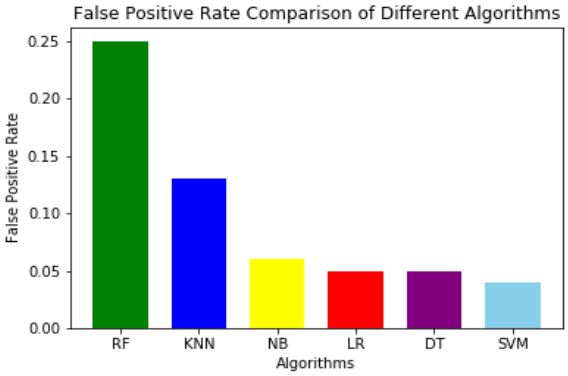

Fig-6: Comparison of false positive rate among all algorithms.

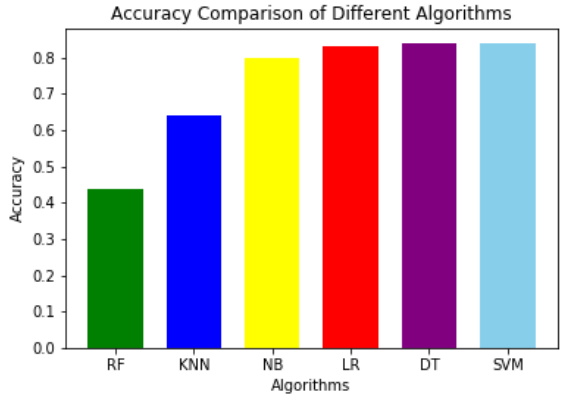

Fig 7: Comparison of accuracy among all algorithms.

From the comparison results shown in Table-7 and corresponding figures, it's clear that Support Vector Machine (SVM) produced best result by outperforming all other algorithms in terms of sensitivity, specificity, precision, false positive rate and accuracy.

V. Conclusion

This paper makes realistic contribution to the field of data science and natural language processing. In this thesis we compare six different classification techniques (Random Forest, K-Nearest Neighbors, Naïve Bayes, Logistic Regression, Decision Tree and Support Vector Machine) with their accuracies. There exist many researches which have been done in the field of sentiment analysis (positive,

negative and neutral) of twitter data but this work focuses on classified twitter data into four broad categories: politics, sports, crime and natural. This system helps users to categories the tweeter posts. In the same time user can understand from this model which type of topic people liked most to share. Among six classification techniques SVM gives more accuracy that is 84%. In future, by increasing the tweeter data and classification category we can build stronger model that gives more accuracy and helps to user preferences.

REFERENCES

[1] Vishal A. Kharde, S.S. Sonawane, Sentiment Analysis of Twitter Data, International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016.

[2] Neha Upadhyay1, Prof. Angad Singh2, Sentiment Analysis on Twitter by using Machine Learning Technique, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue V, May 2016.

[3] Ankita Gupta1, Jyotika Pruthi2, Neha Sahu, Sentiment Analysis of Tweets using Machine Learning Approach, IJCSMC, Vol. 6, Issue. 4, April 2017, pg.444 – 458.

[4] K. Kaviya 1, K.K. Shanthini1, Dr.M. Sujithra2, "Micro-blogging Sentimental Analysis on Twitter Data Using Naïve Bayes Machine Learning Algorithm in Python", International Journal on Future Revolution in Computer Science & Communication Engineering, Volume: 4 Issue: 4, April, 2018.

[5] Bhagyashri Wagh1, J. V. Shinde2, N. R. Wankhade3, Sentiment Analysis on Twitter Data Using Naïve Bayes, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 12, December 2016.

[6] Vikrant Hole1, Mukta Takalikar, Real Time Tweet Summarization and Sentiment Analysis of Game Tournament International Journal of Science and Research (IJSR), 2013.

[7] Bharati S. Kannolli1, Prabhu R. Bevinmarad2 "Analysis and Prediction of Sentiments for Cricket Tweets Using Hadoop", International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 10, oct 2017.

[8] Ankita Rane1, Dr. Anand Kumar2, "Sentiment Classification System of Twitter Data for US Airline Service Analysis", 42nd IEEE International Conference on Computer Software & Applications, 2018.

[9] Nazim Razali1, Aida Mustapha1, Faiz Ahmad Yatim2, Ruhaya Ab Aziz1 "Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)" International Research and Innovation Summit (IRIS 2017).

ABOUT THE AUTHORS

Ananya Sarker is currently working as a Senior Lecturer of Computer Science and Engineering Department, Bangladesh Army University of Engineering & Technology, Bangladesh. She received B.Sc. degree in Computer Science & Engineering from Rajshahi University of Engineering & Technology, Bangladesh in 2013. She is currently pursuing her M.Sc. degree under the supervision of Prof. Dr. Md. Shahid Uz Zaman at the Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology. Her research interests are Machine Learning, Data Mining, Deep Learning, Natural Language Processing and Artificial Intelligence. She is an active member of Institute of Engineers, Bangladesh. She received technical scholarship from Rajshahi University of Engineering & Technology, Bangladesh for brilliant performance in academic career.

Md. Shahid Uz Zaman is a Professor of Computer Science & Engineering Department, Rajshahi University of Engineering & Technology, Bangladesh. He pursued his B.Sc. in Electrical and Electronics Engineering, M.Sc. in Computer Engineering & Ph.D. in Computer and Information Engineering degrees from Rajshahi University of Engineering & Technology, Bangladesh, Shanghai University, China and University of the Ryukyus, Japan respectively. Throughout his career he has published numerous publications. He has three months practical working experience in International System Development (ISD) Company, Okinawa, Japan and working experience with Japan Spatio Temporal Information System (JSTIMS). He worked as a Research Associate for about one year in the department of Information Engineering, University of the Ryukyus, Japan in 2003-2004. His research interests are GIS-based Mapping, VRPs and Satellite Imaging. He is the syndicate member and member of Academic Council, Rajshahi University of Engineering & Technology, Bangladesh. He achieved Chinese Govt. Scholarship by Chinese Govt. in 1994 for Master Program and Japanese Govt. Scholarship by Monbusho in 2000 for Ph.D. Program.

Md. Azmain Yakin Srizon is a Lecturer of Computer Science and Engineering Department, Bangladesh Army University of Engineering & Technology, Bangladesh. He pursued B.Sc. degree in Computer Science & Engineering from Rajshahi University of Engineering & Technology, Bangladesh in 2019. He has previously worked on Prognostic Biomarker Identification of Pancreatic Cancer and published an International Conference Paper on the same topic in 2019. His research interests are Bioinformatics, Machine Learning, Deep Learning, Data Mining, Digital Image Processing, Neural Networks and Artificial Intelligence. In 2018, he became the Champion of Huawei Seeds for the Program 2018 and received a training on Networking, Artificial Intelligence, IoT, Cloud Computing in Huawei Headquarters at Shenzhen, China. He has been certified by both the Huawei University, Shenzhen, China and the Beijing Language and Culture University, Beijing, China. He has been awarded several times throughout his career for punctuality, behavior and discipline.

